Amendments to the Specification:

1. Please perform all of the amendments indicated in the Applicant's **October 20, 2003 response** to the Office Action of April 22, 2003, **except** Items No. 7, 17, 18, 19, 20, 24, and 32, which were objected to in Item 5 in the Examiner's communication of June 21, 2005, and all of which are cancelled below.

2. Please cancel the following Item No. 7 in Applicant's October 20, 2003 response to the Office Action of April 22, 2003:

> 7. Please replace the text beginning on page 9, line 5:
>
> In one embodiment of the invention, the figure-of-merit indices are calculated by obtaining text in the scientific literature about genes on a microarray (using an original method that is part of the invention);
>
> with the following amended text:
>
> In the present invention, clusters of genes are characterized automatically by obtaining text in the scientific literature about genes on a microarray (using an original method that is part of the invention);

3. Please replace the text beginning on page 9, line 5:

> In one embodiment of the invention, the figure-of-merit indices are calculated by obtaining text in the scientific literature about genes on a microarray (using an original method that is part of the invention);

with the following amended text:

> In the invention, text in the scientific literature is obtained about genes on a microarray (using an original method that is part of the invention);

4. Please cancel the following Item No. 17 in Applicant's October 20, 2003 response to the Office Action of April 22, 2003:

> 17. Please replace the text beginning on page 13, line 13:
>
> An objective of the present invention is to generate quantitative indices about
>
> with the following amended text:
>
> An objective of the present invention is to generate words or phrases that describe

5. Please cancel the following Item No. 18 in Applicant's October 20, 2003 response to the Office Action of April 22, 2003:

> 18. Please replace the text beginning on page 13, line 17:
>
> > An advantage of such quantitative indices
>
> with the following amended text:
>
> > An advantage of such descriptors

6. Please delete the following text beginning on page 13, line 13 (which contained the text of the previous two cancellations):

> An objective of the present invention is to generate quantitative indices about the functional, structural, or biochemical pathway relatedness of genes in clusters, using literature databases such as PubMed, in which the connections between genes are implicit in the frequency with which literature about different genes uses the same words and terms. An advantage of such quantitative indices is that, in this invention, they can be generated automatically, and their generation does not force genes into a predetermined classification scheme, which may contain overly-broad or overly-narrow classifications.

7. Please cancel the following Item No. 19 in Applicant's October 20, 2003 response to the Office Action of April 22, 2003:

> 19. Please replace the text beginning on page 13, line 23:
>
> > A yet further advantage of the invention is that it provides an automatic method for identifying the relevant literature for purposes of automatic analysis of the quality of clustering.
>
> with the following amended text:
>
> > A yet further advantage of the invention is that it provides an automatic method for identifying the relevant literature, for purposes of automatic generation of key words or phrases for each cluster.

8. Please replace the text beginning on page 13, line 23:

> A yet further advantage of the invention is that it provides an automatic method for identifying the relevant literature for purposes of automatic analysis of the quality of clustering.

with the following amended text:

> A yet further advantage of the invention is that it provides an automatic method for identifying the relevant literature.

9. Please cancel the following Item No. 20 in Applicant's October 20, 2003 response to the Office Action of April 22, 2003:

20. Please replace the text beginning on page 14, line 2:

A still further advantage of the invention is that for genes associated with a cluster, it provides a ranking of the importance of those genes on the basis of the relevance of text in literature about the set of genes in a cluster. A further advantage of the invention is that it ranks the relatedness of a cluster to all the other clusters, on the basis of the similarity of text in literature about genes in the clusters.

with the following amended text:

A still further advantage of the invention is that for each cluster, the invention ranks key words or phrases according to their importance in characterizing the cluster and distinguishing the cluster from other clusters, in which the ranking is performed after the calculation of a numerical weighting factor.

10. Please delete the text beginning on page 14, line 2:

A still further advantage of the invention is that for genes associated with a cluster, it provides a ranking of the importance of those genes on the basis of the relevance of text in literature about the set of genes in a cluster. A further advantage of the invention is that it ranks the relatedness of a cluster to all the other clusters, on the basis of the similarity of text in literature about genes in the clusters.

11. Please cancel the following Item No. 24 in Applicant's October 20, 2003 response to the Office Action of April 22, 2003:

24. Please replace the text beginning on page 24, line 9:

After downloading all of the Omim Web page files corresponding to the Omim numbers that had been associated with spots on the microarray, the UID Identification module **(118)**

with the following amended text:

After downloading all of the Omim Web page files corresponding to the Omim numbers that had been associated with spots on the microarray, and storing them in the Omim Web Pages section **(146)** of the Data Repository **(138)**, the UID Identification module **(120)**

12. Please cancel the following Item No. 32 in Applicant's October 20, 2003 response to the Office Action of April 22, 2003:

> 32. Please replace the text beginning on page 44, line 19:
>
> > Upon completion of these steps by the Data Summarization Module **(134)**, the Process Control Module **(116)** initiates operation of the Data Output Module **(136)**. It displays the key words or phrases,
>
> with the following amended text:
>
> > Upon completion of these steps, the Process Control Module **(116)** initiates operation of the Data Output Module **(136)**, which displays the key words or phrases that were generated in accordance with the rainbow options that had been selected.

13. Please replace the text beginning on page 44, line 19:

> Upon completion of these steps by the Data Summarization Module **(134)**, the Process Control Module **(116)** initiates operation of the Data Output Module **(136)**. It displays the key words or phrases,

with the following amended text:

> The Process Control Module **(116)** initiates operation of the Data Output Module **(136)**. It displays the key words or phrases.

14. Please perform all of the amendments indicated in the Applicant's **July 07, 2004 response** to the Office Action of January 8, 2004, **except** Items No. 17, 18, and 19, which were objected to in Item 6 in the Examiner's communication of June 21, 2005, and all of which are cancelled below.

15. Please perform all of the amendments indicated in Applicant's **March 21, 2005 response** to the Office communication of February 23, 2005, which supplemented Applicant's July 07, 2004 response to the Office Action of January 8, 2004.

16. Please cancel the following Item No. 17 in Applicant's July 07, 2004 response to the Office Action of January 8, 2004:

> 17. Please replace the text on page 36, line 14:
>
> > It then produces a statistical model of the text that is suitable for text classification.
>
> with the following amended text:
>
> > It then produces a statistical model of the text that is suitable for text classification, although it should be noted that the disclosed invention does not actually perform text classification and does not use the features of Rainbow that actually perform text classification.

17. Please cancel the following Item No. 18 in Applicant's July 07, 2004 response to the Office Action of January 8, 2004:

> 18. Please replace the text on page 38, line 7
>
> The information that is provided automatically by the Keyword Identification Module **(128)** is a list of words for each cluster, sorted in descending order ~~according to the numerical weights calculated by a classification~~ algorithm.
>
> with the following amended text:
>
> The information that is provided automatically by the Keyword Identification Module **(128)** is a list of words for each cluster, sorted in descending order according to the numerical weights calculated by a classification algorithm. (It should be noted that only a portion of the algorithm is used to provide the list of words for each cluster, and that portion does not actually perform text classification. The relevant portion of the algorithm is instead a preliminary or adjunct to the portion that would perform text classification).

18. Please cancel the following Item No. 19 in Applicant's July 07, 2004 response to the Office Action of January 8, 2004:

> 19. Please insert the following text, beginning as a new paragraph, after the text ending on page 10, line 15:
>
> There is little prior art that that can assist an artisan in automatically generating a useful corpus of literature about an individual gene, which might then be used to analyze the literature about the genes that constitute a microarray cluster. PubMed/MEDLINE is the most widely used on-line source for gene related abstracts and literature, which might be used to generate such a corpus, but few investigators have described its use for any similar purpose. SHATKAY et al (2000) explain that PubMed provides for literature search and retrieval by two methods -- boolean query and similarity query (also known as "neighboring"). They describe how there are well-known deficiencies with any attempt to use the method of boolean queries to generate a text corpus. For example, CHAUSSABEL and SHER (2002) attempted to use boolean queries consisting of gene names taken from a list, and ultimately found it necessary to manually edit or correct the unacceptably large number of errors that resulted from use of boolean queries. Accordingly, Shatkay et al. advocate using only the neighboring feature of PubMed to acquire a set of documents about a gene, after first selecting a "kernel" citation for that gene (if possible) within a curated database about the genes under investigation. The literature in PubMed that "neighbors" this kernel citation is then generated by PubMed after providing it the kernel citation as the neighboring query. The method of Shatkay et al then seeks to find similarities within the documents so generated for different genes. This method can be automatic only if there already exists a curated citation list from which to obtain the "kernel" documents, as was the case with the yeast genes investigated by Shatkay et al. Otherwise, and in general, an expert human would need to select the kernel documents. Furthermore, Shatkay et al. teach

that when a clustering of genes is already available from microarray expression experiments, then that clustering should be ignored, except for purposes of manually comparing with results obtained independently by their method. Another method for generating a corpus of text using MEDLINE was described by ANDRADE and VALENCIA (1998), but it was used to generate a corpus only for protein domain families, rather than for individual, arbitrarily selected genes. According to their method, protein families in the PDBSELECT database pointed to entries in the SwissProt database, which pointed to articles in MEDLINE, which were then taken to be the literature corpus for the corresponding protein domain family. This method is not generally applicable to the problem of generating a literature corpus for an arbitrarily selected gene, because a gene may not belong to a known protein family. Furthermore, the size of the literature corpus would be limited by the number of pointers in the SwissProt database. Given the above-mentioned limitations of the prior art, it was therefore an aim of the present invention to provide an original method for automatically generating a substantial literature text corpus for an arbitrarily selected gene, which could be used to generate a literature text corpus for clusters obtained from microarray experiments.